Mathematical Modelling, Vol. 7, pp. 1353–1363, 1986
Printed in the U.S.A.

0270–0255/86 $3.00 + .00

EVALUATING THE EFFECTS OF OBSERVED AND UNOBSERVED DIFFUSION PROCESSES IN SURVIVAL ANALYSIS OF LONGITUDINAL DATA

ANATOLI I. YASHIN*

International Institute for Applied Systems Analysis
and Institute for Control Sciences
Moscow, USSR

KENNETH G. MANTON†

Center for Demographic Studies
Duke University
Durham, NC USA

and

ERIC STALLARD†

Center for Demographic Studies
Duke University
Durham, NC USA

(*Received* 18 *May* 1985)

Abstract—In biostatistical, epidemiological and demographic studies of human survival it is often necessary to consider the dynamics of physiological processes and their influences on observed mortality rates. The parameters of a stochastic covariate process can be estimated using a conditional Gaussian strategy based on the mortality model presented in M. A. Woodbury and K. G. Manton, A random walk model of human mortality and aging. *Theor. Popul. Biol.* **11**, 37–48 (1977) and A. I. Yashin, K. G. Manton, and J. W. Vaupel, Mortality and aging in a heterogeneous population: A stochastic process model with observed and unobserved variables. *Theor. Popul. Biol.*, in press. (1985). The utility of this approach for modeling survival in a longitudinally followed population is discussed—especially in the context of conducing coordinated analyses of multiple similarly constituted databases. Furthermore, the conditional Gaussian approach offers several substantive and computational advantages over the Cameron–Martin approach R. H. Cameron and W. T. Martin, The Wiener measure of Hilbert neighborhoods in the space of real continuous functions. *J. Math. Phys.* **23**, 195–209.

I. INTRODUCTION

In the analysis of mortality and morbidity in demographic, epidemiological and biostatistical studies the explicit relationships between the realized event rates and the parameters of the underlying physiological process generating the health events are often not considered. This can lead to a lack of precision in attempting to use the observed data to make forecasts about health changes, inaccurate statements about the effect on disease risk of altering risk factors in some specified way and an inability to determine the uncertainty of forecasts. In this paper a model is developed which relates the observed mortality rates to the parameters of the underlying physiological processes, both in the

* Postal address: IIASA, Laxenburg, A-2361, Austria.
† Postal address: 2117 Campus Drive, Durham, NC 27706 USA.

presence of auxiliary information on the cross-temporal change of individual physiological values and in the absence of such information. The estimation procedure relies upon the availability of information, either from other studies or from theoretical models, that can be used to specify a reasonable structure for the process. Often, for the analyses of chronic disease such data will be available from prior epidemiological and clinical studies.

This approach is based on a multivariate Gaussian diffusion process model of human physiological change and mortality[1–3]. That model yields the mathematical relationships between the unconditional mortality rates for the population and the parameters of the individual level processes governing change in the means and covariances of the physiological variables related to the risk of mortality. In this paper these relationships are used to compute the unconditional mortality rates for the individual level processes to develop a likelihood function for the estimation of process parameters from the distribution of the observed time to death (failure) of persons in the population. This likelihood function can be used to obtain estimates of process parameters conditional upon the realized values of an observed process measured at fixed times.

The conditional Gaussian approach to estimation employed in this paper can be contrasted with the more usual Cameron–Martin[4] approach (and its extensions[5]) for determining the parameters of a stochastic process. The substantive and computational advantage of the conditional Gaussian approach over Cameron–Martin based strategies will be examined. Furthermore, it will be seen that the model yields, as a special case, strategies for estimating the effects of fixed unobserved covariates on the time of death (e.g. see Ref. [6]) and that the model can be applied to the parallel analyses of survival in multiple data sets where each data set has particular deficiencies.

II. PRELIMINARY SPECIFICATIONS

Suppose that the mortality rate for individual i in a sample of I individuals depends on some l-dimensional physiological process $Z(t)$ which evolves over time. It is assumed that the process for each individual evolves independently from that of all other individuals. The mortality rate for the individual (index i suppressed, except where absolutely essential to convey meaning) is assumed to be a quadratic function of the set of values $Z(t)$, or

$$\mu(t, Z(t)) = Z'(t)Q(t)Z(t) + \mu_0(t), \tag{1}$$

where $Q(t)$ is a non-negative definite symmetric $l \times l$ matrix and $t \geq 0$. The functional form with a linear term on the right-hand side can be easily transformed to (1). The process for the individual $Z(t)$ (index i again suppressed) is assumed to satisfy a linear diffusion type, stochastic differential equation, defined on a probability space (Ω, H, P):

$$dZ(t) = (\alpha_0(t) + \alpha(t)Z(t))\,dt + b(t)\,dW_t, \tag{2}$$

where $\alpha_0(t)$ is a l-dimensional vector function of t with bounded elements for any $t \geq 0$; $\alpha(t)$ is a bounded $l \times l$ matrix for any $t \geq 0$; $b(t)$ is a bounded $l \times k$ matrix, and W_t is a k-dimensional Wiener process which does not depend on the set of initial conditions [i.e. $Z(0)$]. The forms of (1) and (2) were selected because they have been found to adequately describe both risk factor changes and the risk factor dependencies of chronic disease and mortality in a number of longitudinal epidemiological studies[7, 8]. More generally, the form of (1) and (2) will be selected on the basis of prior relevant biostatistical studies or theoretical insights into the physiological processes of interest.

In purely demographic studies of mortality, one often analyzes the unconditional (observed) age specific death rates in the population. More precise evaluation of the mortality

process can be achieved by utilizing the relation between $\mu(t, Z(t))$, the conditional mortality rate at time t for individuals with physiological status $Z(t)$, and $\bar{\mu}(t)$, the unconditional mortality rate at time t, for conditional Gaussian processes of the type described in Woodbury and Manton[1]. Yashin *et al.*[3] present an extension of this model to the case where the mortality rate is influenced by an unobserved process. This relationship, which involves showing how the expectation of the mortality risk can be taken over the process described by (2), can be written symbolically as

$$\bar{\mu}(t) = E(\mu(t, Z(t)) \mid T > t), \tag{3a}$$

where T is the death time of the individual associated with the mortality rate $\mu(t, Z(t))$.

To evaluate this expectation operation one must know the form of the distribution of $Z(t)$ conditional upon survival to time t. In our exposition we will assume that the initial conditions for individual i, $Z(0)$, are random variables with a multivariate normal distribution with the vector of means denoted $m(0)$ and covariance matrix $\gamma(0)$. We will assume that $\mu(t, Z(t))$ is quadratic [see (1)]. With these assumptions the unconditional mortality rate at time t, $\bar{\mu}(t)$, has the following relation to the parameters of the distribution of $Z(t)$:

$$\bar{\mu}(t) = m'(t)Q(t)m(t) + \mathrm{Tr}(Q(t)\gamma(t)) + \mu_0(t). \tag{3b}$$

If the observations are i.i.d., then $\bar{\mu}(t)$ in (3b) can be interpreted in demographic terms as the cohort mortality rate among survivors to age t.

Assuming (1) and (2) one can show[9] that the changes in the mean vector and covariance matrix satisfy the nonlinear ordinary differential equations,

$$\frac{dm(t)}{dt} = \alpha_0(t) + \alpha(t)m(t) - 2\gamma(t)Q(t)\, m(t) \tag{4}$$

and

$$\frac{d\gamma(t)}{dt} = \alpha(t)\gamma(t) + \gamma(t)\alpha'(t) + b(t)b'(t) - 2\gamma(t)Q(t)\, \gamma(t) \tag{5}$$

with the initial state described by $m(0)$ and $\gamma(0)$.

The relationships specified in (3)–(5) can be used in developing both a substantively meaningful model of human mortality and a statistical estimation procedure for evaluating the parameters of the process describing the evolution of influential factors $Z(t)$ [Eq. (2)], and of the risk function [Eq. (1)]. Specifically, these relationships can be used to integrate ancillary information and data from empirical studies or theoretical insights into a probabilistic model of human failure processes. For example, prior studies may help in selecting the appropriate functional form for $Q(t)$, by providing the form of the functional dependency of the population hazard rate on the means of $Z(t)$ (e.g. see Ref. [10]). Additionally, one may have information on the form of the process $Z(t)$ influencing mortality. Such ancillary information on the structure of the model can increase the precision of forecasts of population hazard rates over naive procedures which ignore this information, e.g. procedures which simply extrapolate the temporal trends of age-specific mortality rates.

In addition to helping to organize ancillary information in the development of a probabilistic model of human failure processes, the relationships in (3)–(5) can be used to obtain maximum likelihood estimates of the parameters of the process from the observed death times in a population and from available longitudinal information on physiological factors. The development of such estimates can be considered in the context of two distinct

observational plans. The first plan is for the continuous time monitoring of mortality, where mortality is influenced by both observed and unobserved processes. Such an observational plan, however, is seldom found in longitudinal epidemiological studies. Furthermore, because the continuous time formulation requires estimating parameters over the entire process from birth to death, it is computationally difficult. As a consequence this type of observational plan is primarily of theoretical interest in dealing with such problems as developing estimation strategies for unequal follow-up times or in developing and assessing interval approximation formulas. More practically, for empirical applications an approach is presented for a second type of observational plan where measurements are made periodically at fixed (i.e. nonrandom) discrete times. The equations developed for the continuous time case for the conditional Gaussian model can be employed in obtaining the maximum likelihood estimates of process parameters from the discrete time measurements if the correct initial conditions for the start of each interval are formulated. This type of observational plan is found in many epidemiological studies (e.g. Framingham, Massachusetts; Evans County, Georgia).

III. A MODEL FOR ESTIMATING THE PARAMETERS OF A TWO-COMPONENT FAILURE PROCESS UNDER BOTH CONTINUOUS AND DISCRETE TIME OBSERVATIONAL PLANS

The first step in the development is to generalize the mortality process defined in (1) and (2) to the case where mortality is influenced by both an observed and unobserved process. Specifically, suppose that the duration of life for any individual in the cohort is a functional of the two component processes $Z'(t) = (Y'(t), X'(t))$. The quadratic form in (1) may then be rewritten as

$$\mu(t, X(t), Y(t)) = (Y'(t), X'(t)) \begin{bmatrix} Q_{11}(t), & Q_{12}(t) \\ Q_{21}(t), & Q_{22}(t) \end{bmatrix} \begin{bmatrix} Y(t) \\ X(t) \end{bmatrix} + \mu_0(t), \tag{6}$$

where $Q_{11}(t)$ and $Q_{22}(t)$ are positive definite symmetric matrices, and $Q'_{12}(t) = Q_{21}(t)$. Furthermore, (2) becomes

$$\mathrm{d}\begin{bmatrix} Y(t) \\ X(t) \end{bmatrix} = \left[\begin{bmatrix} \alpha_{01}(t) \\ \alpha_{02}(t) \end{bmatrix} + \begin{bmatrix} \alpha_{11}(t), & \alpha_{12}(t) \\ \alpha_{21}(t), & \alpha_{22}(t) \end{bmatrix} \begin{bmatrix} Y(t) \\ X(t) \end{bmatrix} \right] \mathrm{d}t + \begin{bmatrix} b_1(t) \\ b_2(t) \end{bmatrix} \mathrm{d}\begin{bmatrix} W_{1t} \\ W_{2t} \end{bmatrix}, \tag{7}$$

where W_{1t} and W_{2t} are vector-valued Wiener processes, independent of each other and of the initial values $X(0)$ and $Y(0)$; $b_1(t)$ and $b_2(t)$ are matrices with the appropriate dimensions. Thus, the processes $X(t)$ and $Y(t)$ are the solution of these linear stochastic differential equations. Let us now consider the two different observational plans for multicomponent processes of the type describe by (6) and (7).

A. *Continuous observations*

Yashin *et al.*[3] gave a solution of (6) and (7) in the conditional Gaussian case by assuming that the distribution of the $Y(t)$ was normal conditional on the observed process. The validity of this assumption follows from the normality of $Z(t)$ because one can always find a vector function $F(t, X(t))$ and a scalar $G(t, X(t))$ such that the individual mortality rate, $\mu(t, X(t), Y(t))$, can be written

$$\mu(t, X(t), Y(t)) = (Y(t) - F(t, X(t))' Q_{22}(t) (Y(t) - F(t, X(t)) + G(t, X(t)), \tag{8}$$

where

$$F(t, X(t)) = -Q_{11}^{-1}(t)\, Q_{12}(t)\, X(t) \tag{9}$$

and

$$G(t, X(t)) = X'(t)\, Q_{22}(t)\, X(t) - X'(t)\, Q_{21}(t)\, Q_{11}^{-1}(t)\, Q_{12}(t)\, X(t) + \mu_0(t). \tag{10}$$

The structure of (8) with respect to $Y(t)$ is similar to the hazard function considered by Myers[5]. However, because of boundary conditions it is difficult to use additional observations on the measured process $X(t)$ in his formulation[9]. A more appropriate strategy seems to involve use of the conditional Gaussian approach developed in Yashin *et al.*[3] for a continuously observed process. This latter approach can be used for the evaluation of a process that is still under observation, e.g. to analyze data from the intermediate phases of a longitudinal study.

B. *Fixed time observation*

Let us now assume that the elements of $X(t)$ are measured at a set of fixed times. Thus $X_i(t_i), \ldots, X_K(t_k)$ are the measurements on the ith individual. $Y_i(t)$ represents the variables that are not measured. Suppose that both processes influence the mortality rate and that this dependence is as described by (6). Furthermore, suppose that the evolution of $X(t)$ and $Y(t)$ are described by (7). Our goal is to estimate the elements of $Q(t)$ in (6) on the basis of data only on X, i.e. $X_i(t_j)$, $i = 1, \ldots, I$, $t_j < T_i$, where T_i is the observed death time for individual i. For simplicity, let the index i be suppressed and $\hat{X}(t)$ be the matrix $X(t_1), X(t_2), \ldots, X(t_j(t))$, where

$$t_j(t) = \sup\{t_j : t_j < t\}. \tag{11}$$

The survival function, conditional on the observed process $\hat{X}(t)$, say $S(t, \hat{X}(t))$, may be defined

$$S(t, \hat{X}(t)) = P(T > t \mid \hat{X}(t)) \tag{12}$$

so that

$$\mu^*(t, \hat{X}(t)) = -\frac{\partial}{\partial t} \ln S(t, \hat{X}(t)), \tag{13}$$

where $\mu^*(t, \hat{X}(t))$ is the mortality rate implied by the conditional survival function. This requires the assumption that the conditional survival function is absolutely continuous. In order to develop an estimation strategy the relation of $\mu^*(t, \hat{X}(t))$ to the parameters of the underlying process and measurements must be found. The appropriate relations are presented in terms of the means $m(t)$, and conditional covariances $\gamma(t)$, of the variables in both X and Y in the following theorem:

THEOREM. Suppose that the process is defined by both measured and unmeasured variables with the structure presented in (7). Then the mortality rate $\mu^*(t, \hat{X}(t))$ can be represented as follows:

$$\mu^*(t, \hat{X}(t)) = m'(t, \hat{X}(t))\, Q(t)\, m(t, \hat{X}(t)) + \mathrm{Tr}(Q(t)\, \gamma(t)) + \mu_0(t), \tag{14}$$

where

$$m(t) = \begin{bmatrix} m_1(t) \\ m_2(t) \end{bmatrix}, \qquad \gamma(t) = \begin{bmatrix} \gamma_{11}(t), & \gamma_{12}(t) \\ \gamma_{21}(t), & \gamma_{22}(t) \end{bmatrix}$$

on the intervals $t_j \leq t < t_{j+1}$ satisfy the equations,

$$\frac{dm(t)}{dt} = \alpha_0(t) + \alpha(t)\, m(t) - 2\gamma(t)\, Q(t)\, m(t) \tag{15}$$

and

$$\frac{d\gamma(t)}{dt} = \alpha(t)\, \gamma(t) + \gamma(t)\alpha'(t) + b(t)b'(t) - 2\gamma(t)\, Q(t)\, \gamma(t) \tag{16}$$

where

$$\alpha_0(t) = \begin{bmatrix} \alpha_{01}(t) \\ \alpha_{02}(t) \end{bmatrix} \quad \text{and} \quad \alpha(t) = \begin{bmatrix} \alpha_{11}(t), & \alpha_{12}(t) \\ \alpha_{21}(t), & \alpha_{22}(t) \end{bmatrix}.$$

The theorem demonstrates that the general form of the hazard rate in (3b) for multivariate normal physiological variables applies directly to the case with fixed measurement times and unobserved and partly observed influential processes. The primary difference between these equations and those for the continuous time case is that a new set of initial conditions holds for each interval $[t_j, t_{j+1})$. This defines an observational plan where, at each time of measurement, there is a jump in information on the observed process and on mortality status. Specifically, at time $t_j, j = 1, \ldots, K$ the initial values for the equation are

$$m_1(t_j) = m_1(t_{\bar{j}}) + \gamma_{12}(t_{\bar{j}})\, \gamma_{\bar{2}\bar{2}}^{-1}(t_{\bar{j}})\, (X(t_{\bar{j}}) - m_2(t_{\bar{j}})), \tag{17}$$

$$m_2(t_j) = X(t_j), \tag{18}$$

$$\gamma_{11}(t_j) = \gamma_{11}(t_{\bar{j}}) - \gamma_{12}(t_{\bar{j}})\, \gamma_{\bar{2}\bar{2}}^{-1}(t_{\bar{j}})\, \gamma_{21}(t_{\bar{j}}), \tag{19}$$

$$\gamma_{22}(t_j) = 0, \tag{20}$$

$$\gamma_{12}(t_j) = \gamma_{21}(t_j) = 0, \tag{21}$$

where $t_{\bar{j}}$ represents the left-handed value of the process, i.e. at the point just before the jump in information. Equations (17) and (18) show how the results of measurements on the process $X(t)$ at each time t_j can be introduced into (14). Thus the mean for $X(t_j)$ is equal to the observed value at the time of measurement (18) while the mean for $Y(t_j)$ is the mean conditional on $X(t_j)$. The variances of the values of the observed variables are equal to 0 at the measurement time; for $Y(t_j)$, the variances are conditional on the values of the observed variables. The initial conditions for each interval represent the jumps in information at these points. Initialization of $m_1(t_1)$ and $\gamma_{11}(t_1)$ for the first interval may require auxiliary data.

The theorem [i.e. Eqs. (14)–(21)] follows from the Kolmogorov–Fokker–Planck (KFP) equation for the special case of the multivariate normal distribution with quadratic hazard and linear dynamics. The proof requires showing that the changes in the multivariate normal distribution described by the KFP equation, conditional upon the probability of

survival *and* the effects of an unobserved Gaussian process, must produce a multivariate normal distribution at each point in time. This can be demonstrated in two stages. The first and most important stage is to prove the conditional Gaussian property. This is done by examining the characteristic function conditional on the process $X(t)$ and survival at least to time t. Once the conditional Gaussian properties are demonstrated (for details see Yashin *et al.*[3]), only the means and variances of the distribution of $Y(t)$ are required to characterize the process. The second stage is to specify the equations for the means and variances, again from the characteristic function.

IV. ESTIMATION

Statistical analyses of the problem can be conducted by the maximum likelihood approach. In order to do that, one needs to specify the parametric uncertainties in the likelihood function. Generally speaking the sources of uncertainties can be the functions $\alpha_0(t)$, $\alpha(t)$, $b(t)$, $Q(t)$, and $\mu_0(t)$. Assume that all of these functions can be written in some parametric form: $\alpha_0(\beta, t)$, $\alpha(\beta, t)$, $b(\beta, t)$, $Q(\beta, t)$, $\mu_0(\beta, t)$. The simplest case of this form is that the coefficients α_0, α, b, Q, and μ_0 are constants. In this case the components of β correspond to the respective coefficients.

The likelihood function for a sample of I persons is

$$\mathscr{L} = \prod_{i=1}^{I} \mu^*(T_i, \hat{X}_i(T_i)) \exp\left\{ -\int_0^{T_i} \mu^*(u, \hat{X}_i(u))\, du \right\} \prod_{j=2}^{K_i} f[X_i(t_j) \mid \hat{X}_i(t_{j-1})], \quad (22a)$$

where $f[X_i(t_j) \mid \hat{X}_i(t_{j-1})]$ is the n-variate conditional Gaussian density of $X_i(t_j)$ given the prior observations in $\hat{X}_i(t_{j-1})$; alternatively, to be consistent with standard notation, one can write this density as $N[X_i(t_j) \mid m_{i2}(t_{\bar{j}}), \gamma_{i22}(t_{\bar{j}})]$, where $m_{i2}(t_{\bar{j}})$ and $\gamma_{i22}(t_{\bar{j}})$ are the (conditional) means and variances exhibited in (17) and (19), respectively. Equation (14) shows how $\mu^*(t, \hat{X}_i(t))$ depends on the means and variances of the influential process. Equations (15) and (16) describe the change with time of the means and variances by a process with coefficients $\alpha_0(t)$, $\alpha(t)$, $b(t)$, and $Q(t)$. These equations permit (22a) to be revised as

$$\begin{aligned}\mathscr{L} = \prod_{i=1}^{I} & \mu^*(T_i, m_i(T_i, \beta, \hat{X}_i(T_i)), \gamma_i(T_i, \beta), \mu_0(\beta, T_i), Q(\beta, T_i)) \\ & \times \exp\left\{ -\int_0^{T_i} \mu^*(u, m_i(u, \beta, \hat{X}_i(u)), \gamma_i(u, \beta), \mu_0(\beta, u), Q(\beta, u))\, du \right\} \\ & \times \prod_{j=2}^{K_i} (2\pi)^{-n/2} \mid \gamma_{i22}(t_{\bar{j}}, \beta) \mid^{-1/2} \exp\{-\tfrac{1}{2}[X_i(t_j) - m_{i2}(t_{\bar{j}}, \beta, \hat{X}_i(t_{\bar{j}}))]' \\ & \times \gamma_{i22}^{-1}(t_{\bar{j}}, \beta)\, [X_i(t_j) - m_{i2}(t_{\bar{j}}, \beta, \hat{X}_i(t_{\bar{j}}))]\}. \end{aligned} \quad (22b)$$

Equations (22a) and (22b) require that the time to failure (T_i) is known for each person in the population. In many practical examples, the time to failure is not observed (i.e. the study is terminated before all persons have experienced the event of interest). In these cases the likelihood must be adjusted to reflect this right censoring. This is accomplished by introducing into Eqs. (22a and 22b) terms describing the probability of survival to the end of the study. For the survivors themselves, only the coefficients of the dynamic equations [i.e. Eq. (7)] can be estimated. To estimate both the dynamics and mortality coefficients, we require the joint likelihood for the subset S of survivors to the end of the study (at time $C_i = C$), and for the subset $\bar{S}$ of nonsurvivors. This likelihood may be

written

$$\mathcal{L} = \prod_{i=1}^{I} [\mu^*(T_i, \hat{X}_i(T_i))]^{\delta_i} \exp\left\{ - \int_0^{T_i \wedge C} \mu^*(u, \hat{X}_i(u))\, du \right\} \prod_{j=2}^{K_i} N[X_i(t_j) \mid m_{i2}(t_j), \gamma_{i22}(t_j)], \tag{23}$$

where $\delta_i = I(T_i \leq C)$ is an indicator variable denoting whether ($\delta_i = 1$) or not ($\delta_i = 0$) the observation is censored. We see that the primary difference between (22a), (22b) and (23) is the inclusion of terms describing the probability of surviving to the end of the study for each member of the surviving subset. More general types of censoring can be handled if C in (23) can be replaced with the exact time C_i of censoring on an individual basis.

To obtain maximum likelihood estimates of the process parameters in (22a), (22b) or (23), one will need appropriate numerical procedures. The details of these procedures will depend on the specific assumptions made about the characteristics of both the unobserved and the observed covariate processes. For example, to evaluate $\mu_i^*(\cdot)$ in (14) at the time of measurement for an observed covariate process, one needs values for $Q(t)$ and $\mu_0(t)$, but can set the means equal to the observed values [see (18)] and the variances to zero [see (20)]. To calculate the hazard rate for any intermediate time between measurements however, one needs values of $\alpha_0(t)$, $\alpha(t)$, $b(t)$, and $Q(t)$ to solve (15) and (16) so that the projected values of $m(t)$ and $\gamma(t)$ may be substituted in (14). Thus the mortality parameters, $Q(t)$ and $\mu_0(t)$, and the covariate process parameters, $\alpha_0(t)$, $\alpha(t)$, and $b(t)$, are inextricably intertwined. One solution suggested by Myers[5] is to measure the covariates sufficiently often that the covariate path is effectively known. Implementation of this strategy for the special case where the subject intervals are equal for all subjects and for all measurement times is described in Manton *et al.*[11].

Estimation in the case of unobserved covariate processes will depend on insights available from relevant biomedical theory and from related auxiliary data. For example, in the case of multiple data sets with different sets of measurements, certain variables may be available in one data set but not in others. In this case, from the data set where a variable is measured we can obtain direct estimates of the corresponding coefficients in $\alpha_0(t)$, $\alpha(t)$, $b(t)$, $Q(t)$, $m(t)$, and $\gamma(t)$. This suggests that estimation of parameters for all data sets be conducted jointly with the sets of X and Y variables being redefined from one data set to the next. Where there are Y variables not measured in any of the data sets[6], one can still implement the model if there is sufficient theoretical evidence to specify the form of the initial distributions, the dynamics, and the mortality risks associated with the unobserved process (see Yashin *et al.*[3] for discussion).

V. A COMPARISON OF THE CAMERON–MARTIN AND CONDITIONAL GAUSSIAN APPROACHES

The Cameron–Martin approach[9] gives a way of calculating the mathematical expectation of an exponent which is a functional of a Wiener process. The exponent can be considered as a conditional survival function. Thus the approach has been suggested as a methodology for survival analysis where the stochastic process in the exponent is interpreted as covariates affecting the survival rate. Unfortunately, the Cameron–Martin approach has several significant limitations. To illustrate for a linear diffusion process of the type in (7) but with $\alpha_0(t) \equiv 0$ (i.e. no "drift"), the matrix of hazard coefficients $Q(t)$, has the property

$$E \exp\left\{-\int_0^t \begin{bmatrix} Y(u) \\ X(u) \end{bmatrix}' Q(u) \begin{bmatrix} Y(u) \\ X(u) \end{bmatrix} du\right\}$$

$$= \exp\left\{\begin{bmatrix} Y(0) \\ X(0) \end{bmatrix}' \Gamma(0) \begin{bmatrix} Y(0) \\ X(0) \end{bmatrix} + \operatorname{Tr} \int_0^t \begin{bmatrix} b_1(u) \\ b_2(u) \end{bmatrix} \begin{bmatrix} b_1(u) \\ b_2(u) \end{bmatrix}' \Gamma(u)\, du\right\}, \quad (24)$$

where $\Gamma(u)$ is the solution of the matrix Ricatti equation

$$\frac{d\Gamma(u)}{du} = Q(u) - (\Gamma(u) + \Gamma'(u))\alpha(u) - \frac{1}{2}(\Gamma(u) + \Gamma'(u)) \begin{bmatrix} b_1(u) \\ b_2(u) \end{bmatrix} \begin{bmatrix} b_1(u) \\ b_2(u) \end{bmatrix}' (\Gamma(u) + \Gamma'(u)) \quad (25)$$

with the terminal condition $\Gamma(t) = 0$.

The particular case of formula (24) corresponds to the well-known Cameron–Martin results[9] specified for a Wiener process in the exponent of the form:

$$E \exp\left[-\int_0^t (W_u, Q(u)W_u)\, du\right] = \exp\left[\frac{1}{2}\int_0^t \operatorname{Tr} \Gamma(u)\, du\right], \quad (26)$$

where $(W_u, Q(u)W_u)$ is the scalar product equal to the quadratic form, $W_u' Q(u)W_u$, and $\Gamma(u)$ is a symmetric nonpositive definite matrix which is the unique solution of the matrix Riccati equation

$$\frac{d\Gamma(u)}{du} = 2Q(u) - \Gamma^2(u) \quad (27)$$

and $\Gamma(t) = 0$ is a zero matrix.

To prove these relations one uses likelihood ratio principles applied to diffusion-type processes[12, 13]. Using this approach, Myers[5] found the formulas for averaging the exponent when, instead of a Wiener process, there is a process satisfying a linear stochastic differential equation driven by a Wiener process [i.e. (24) and (25)].

Unfortunately, the proof of the Cameron–Martin formula and its generalization[5] does not use the interpretation of the matrix Q as hazard coefficients and does not provide a direct physical interpretation of the variables $\Gamma(u)$ in (24) [or (26)]. Furthermore, the boundary condition [i.e. $\Gamma(t) = 0$] on (25) [and (27)] makes it difficult to conduct the calculations either for subintervals or for extended intervals when additional longitudinal measurements are made.

The methods described in this paper do not have these limitations. They involve the use of "martingale" techniques to produce a general formula for averaging exponents which can be a more complex functional of a random process of a wider class[9]. This paper provides the specialization of these procedures to the case where the functional is a quadratic form for averaging the exponents.

VI. DISCUSSION

We presented a procedure for evaluating the stochastic process underlying the observed population averaged survival rate. This procedure, using conditional Gaussian properties, can lead to computationally powerful likelihood ratio techniques for assessing human survival data which have superior properties to the Cameron–Martin procedure. Specif-

ically, they do not have the limitation of the Cameron–Martin approach that, due to its boundary conditions, all parameters of the process must be recalculated when the study period is extended.

The procedure is applicable in several important areas. First, there has been much recent attention to the question of heterogeneity (unmeasured differentials in transition rates) and its effects on the analysis of human survival[14, 15, 6]. Underlying this concern is the analytic problem of how systematic selection of persons by mortality affects the average force of transition among survivors. This involves examination of the effects of averaging the exponent (and related functional) in the survival function. Past efforts to resolve the problem in the analysis of human survival have been to ignore the effects of diffusion by assuming a deterministic trajectory for the temporal dependence of the individual hazard rate. This approximation can be problematic in certain applications, especially in attempts to infer the operation of the risk mechanism in elderly individuals, where the forces of homeostasis may be weakening[10]. By explictly including the diffusion process in the proposed model one can potentially greatly improve the precision of model-based predictions and certainly have a better procedure for determining the effects of intervention on the realization of risk.

A second major utility is that the proposed approach facilitates the introduction of auxiliary information into analyses of the failure process. This is because one can directly specify the details of the process and thereby introduce information into the appropriate features of the model. This is critically important in analyzing human survival at advanced ages because the evolution of chronic diseases is a complex process operating over a lengthy time scale. Thus, though there is considerable empirical information from existing longitudinal studies on risk covariates and on the evolution of chronic disease, seldom have the dynamic properties of such data been completely exploited. For example, certain negative associations have been demonstrated between risk factors (e.g. asbestos) and specific disease outcomes (e.g. lung cancer) because of the systematic selection of susceptible persons by disease processes (e.g. asbestosis) which had an earlier age assault pattern[16]. Such dynamics and systematic selection require consideration of the basic dynamic process and the effects of selection on the average risk among survivors to unconfound such factors. Only by using auxiliary information and a model of the intrinsic processes can such public health questions be adequately resolved.

An additional important area of application in health studies is in developing procedures applicable to the coordinated analyses of multiple data sets. Specifically, we often find that any single longitudinally followed study population will have important limitations. For example, such study populations are usually of limited size so that there is often inadequate information to assess the risks of specific disease outcomes. Second, in order to increase the efficiency of data collection the study population is often deliberately truncated so that only age groups with significant event rates (but usually low rates of loss) are followed. Finally, different sets of measurements are made in the various studies. The conditional Gaussian strategies presented in this paper provide means by which information can be combined across studies (by being built into parameters of the underlying process) and by which the parameters of the process can be estimated from segmented (and hence truncated) data.

Acknowledgment—The efforts of K.G.M. and E.S. in this research were supported by NIA Grant No. AG01159-09 and NSF Grant No. SES8219315.

REFERENCES

1. M. A. Woodbury and K. G. Manton, A random walk model of human mortality and aging. *Theor. Popul. Biol.* **11**, 37–48 (1977).

2. M. A. Woodbury and K. G. Manton, A mathematical model of the physiological dynamics of aging and correlated mortality selection: Part I—Theoretical development and critiques. *J. Gerontol.* **38**, 398–405 (1983).
3. A. I. Yashin, K. G. Manton and J. W. Vaupel, Mortality and aging in a heterogeneous population: A stochastic process model with observed and unobserved variables. *Theor. Popul. Biol.*, in press (1985).
4. R. H. Cameron and W. T. Martin, The Wiener measure of Hilbert neighborhoods in the space of real continuous functions. *J. Math. Phys.* **23**, 195–209 (1944).
5. L. Myers, Survival functions induced by stochastic covariate processes. *J. Appl. Prob.* **18**, 523–529 (1981).
6. J. J. Heckman and B. Singer, Population heterogeneity in demographic models. In *Multidimensional Mathematical Demography*, K. Land and A. Rogers, eds., pp. 567–599, Academic Press, New York (1982).
7. K. G. Manton and M. A. Woodbury, A mathematical model of the physiological dynamics of aging and correlated selection: Part II—Application to the Duke Longitudinal Study. *J. Gerontol.* **38**, 406–413 (1983).
8. K. G. Manton and M. A. Woodbury, A continuous-time multivariate Gaussian stochastic process model of change in discrete and continuous state variables. In *Sociological Methodology, 1985* N. Tuma, ed., Jossey-Bass, in press (1985).
9. A. I. Yashin, Dynamics in survival analysis: Conditional Gaussian property versus Cameron–Martin formula. WP-84, International Institute for Applied Systems Analysis, Laxenburg, Austria (1984).
10. A. C. Economos, Rate of aging, rate of dying and the mechanism of mortality. *Arch. Gerontol. Geriatr.* **1**, 3–27 (1982).
11. K. G. Manton, E. Stallard and M. A. Woodbury, Chronic disease evolution and human aging: A general model for assessing the impact of chronic disease in human populations. *Int. J. Math. Model* in press (1985).
12. A. A. Novikov, On parameters estimation of diffusion processes. *Studia Sci. Math.* **7**, 201–209 (1972).
13. R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes*, Springer-Verlag, New York (1977).
14. J. W. Vaupel, K. G. Manton and E. Stallard, The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demogr.* **16**, 439–454 (1979).
15. K. G. Manton and E. Stallard, Heterogeneity and its effect on mortality measurements. Chapter 12 in *Methodologies for the Collection and Analysis of Mortality Data*, J. Vallin, H. Pollard and L. Heligman, eds., pp. 265–299, IUSSP, Ordina Editions (1984).
16. K. G. Manton, An evaluation of strategies for forecasting the implications of occupational exposure to asbestos. U.S. Library of Congress, Congressional Research Service, Government Division (1985).